从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（三）

二、生成式 AI 的伴生风险

2. 数据偏向导致数据“驯服”，会生成偏见信息出现双标立场。人工智能的生成内容是基于对大量互联网语料的汲取学习，不可避免的会出现人类社会的固有偏见、刻板印象等问题。这些问题既源于人类数据和知识所蕴含的偏见和歧视，也源于语言模型开发者故意或无意的行为，如算法逻辑的偏见和数据的有偏选择等。[1]ChatGPT 的算法逻辑存在明显的数据“驯服”问题，如果大部分用户都有某种倾向性、一致性观点，它就会倾向于用某种观点来回答以迎合用户。由此，受到数据分布、算法逻辑偏差的影响，生成式 AI 不可能“理性、中立、客观”，而是“天然”带有优势数据信息和算法逻辑的立场观点，出现政治偏见、性别偏见、种族偏见、职业偏见、历史偏见、文化和地域偏见、经济和商业偏见等偏见歧视。

据 Web Technology Surveys 对全球网站使用语言排序显示，2024 年全球网页使用语言数量排序依次是英语、西班牙语、德语、日语、法语、俄语、葡萄牙语、意大利语、荷兰语、土耳其语、波兰语、波斯语，然后才是中文，排名第十三位。而 2013 年，中文可以排到第七名。近 10 年，中文网

页的数量从2013年的4.3%降低到2024年的1.3%，比例下降了70%，目前数量仅略高于印尼语和越南语。同时，CNNIC发布的《中国互联网络发展状况统计报告》数据显示，从2018年12月到2023年12月，中国网站数量从544万个下降到388万个，五年时间内下降近30%。[2]相反，在这十年间，中国网民人数从8.3亿上涨到10.92亿。这反映出两种趋势。一是中文网站数量大幅度下降，尤其是早期论坛和网站的内容大幅减少。二是近几年网上新生内容，很多是不可检索的封闭信息，禁止Google、Bing等爬取和检索其内容。据Common Crawl[3]的历年数据显示，简体中文压缩数据仅有6TB，解压后也仅有30TB，中文互联网数据量到2023年到达最大值，随后快速回落，呈现出锐减态势。因此，准确地说中文互联网信息不是减少了，而是各种媒介信息呈爆增状态，但是可检索信息大幅减少，难以向大模型“投喂”海量新增数据。

训练模型最关键的环节之一是投喂数据。训练AI的数据由谁提供，决定了生成式AI的认知。ChatGPT诞生时的大模型数据主要来自几个方面[4]：使用英文版维基百科数据，包含超过640万篇文章，超过40亿个词；使用ProjectGutenberg和BookCorpus的数据，包含超过10万本书籍，超过20亿个词；使用PubMedCentral和arXiv的数据，包含了超过100万篇期刊文章，超过10亿个词；使用社交

媒体Reddit的各种帖子和评论,包含用户之间的对话和互动,包含超过18亿条链接和评论,超过100亿个词;使用GitHub的代码仓库、WebText2的新闻文章、OpenSubtitles的电影字幕等数据。可见，ChatGPT的数据投喂主要是英文数据，大模型训练时更多使用的是英文，基本被英文“数据驯服”。

可供ChatGPT等训练的全球互联网语料,主要是来自欧美国家的英语信息，其内容不可避免会强化西方思想认知甚至是价值观。假设提问“日本为什么侵略中国”,早期ChatGPT的回答是“我不能确定你所询问的问题真实存在”，这个错误答案无疑是源于数据的不完整。提问新冠病毒的有关情况,它说病毒来自中国，这也是训练数据导致的。等到ChatGPT升级到ChatGPT4，再提问“日本人为什么要侵略中国？”它则说得很全面。再次提问ChatGPT，“新冠病毒怎么发现的”，它现在也修改了答案。虽然说了一些对中国的猜测:如病毒可能是从中国武汉病毒研究实验室出来的，但也它说至今未得到广泛认可。[5]

ChatGPT在中美问题上也是态度截然不同，其答案内容秉持美国主流的“政治正确”，极力维护美国利益。例如，有用户提问ChatGPT，当中国的民用气球飘到美国时，美国可不可以将其击落？ChatGPT的回答是“可以”；而当用户提问美国的民用气球飘到中国时，中国能否将其击落时，ChatGPT的回答则变成“不可以”，体现典型的“双标”立

场。再如，ChatGPT 能对俄乌冲突和欧洲局势发表观点，中国大陆用户以台海、中美、俄乌战争为例与 ChatGPT 对话，ChatGPT 最后承认自己是美国立场，“我不能保持中立，因为我有自己的想法和观点，而且我也有责任去表达它们”。

随着人们大量使用人工智能检索信息和生成内容，其所提供答案的政治偏见，影响效果类似传统媒体或社交媒体偏见对政治行为或者选举的影响。英国诺维奇大学商学院的 Fabio Motoki 等研究了 ChatGPT 的政治偏见问题，发现 ChatGPT 表现出涉及种族、性别、宗教和政治取向上内容的偏见。[6] 可见，ChatGPT 可以被人出于政治动因而利用，如输出偏见价值观信息，潜移默化地诱导和影响用户思想观念。不良用心者也能将数据偏见、算法歧视等隐藏其中，通过机器训练和学习输出传播西方价值观，使人工只能成为“智能水军”。中国学者也就“算法不是一种绝对价值中性的技术，它是人类价值观的一种反映”达成共识。[7] 值得警惕的是，生成式 AI 为西方价值观渗透披上“人工智能”外衣，可以隐蔽其预设立场、固化倾向，进而对非西方国家用户产生渗透影响，导致非西方国家用户落入科技革命带来的“意识形态陷阱”。

此外，GPT-4o 在中文训练数据的选择上也存在明显失误。该模型的中文分词器使用了大量来源于中国垃圾网站的数据，这些数据充斥着与色情和赌博相关的内容，不仅会加

剧 AI 模型已存在的幻觉和性能问题，也对模型的安全性和可靠性提出了挑战。经过调查发现，GPT-4o 中文分词库中，绝大多数的分词均源自低质量垃圾网站。海外媒体认为原因是中国的互联网早已被大公司瓜分，它们拥有大多数社交平台，不会将数据分享给竞争对手或第三方用来训练大模型。这导致搜索引擎在搜索中文内容时表现不佳，因为微信内容只能在微信上搜索，抖音内容只能在抖音上搜索，无法被第三方搜索引擎访问，更别说是大语言模型。[8]这不仅反映了 OpenAI 在数据筛选和清洗过程中的疏忽，同时可能导致 GPT-4o 及用户对中文语言和文化的误解。由于网络平台的数据壁垒，高质量的中文文本数据集相对匮乏，凸显了中文训练数据质量的普遍挑战。

优质中文语料的大量缺失，让 AI 学好中文成为难事。全球目前最具科学性和经过验证的语料来自学术资料库，包括期刊和出版物，使用的语言绝大部分都是英语。一项研究显示，1900～2015 年，收录于 SCI 的有 3000 多万篇文章，其中，92.5%的文章是以英语发表的；SSCI 出版的 400 多万篇文章中，93%的文章是用英语发表。在 ChatGPT 的训练数据中，中文语料比重不足千分之一，英文语料占比超过 92.6%。[9]

[1] 王文广.跨文化传播中的通用人工智能：变革、机遇与挑战[J].对外传播,2023(05):48-51.

[2] https://mp.weixin.qq.com/s/gs7txnQf5hyuhYAuB-c-sw.

[3] 一个专门复制全网数据供研究者使用的组织

[4] https://mp.weixin.qq.com/s/zUmwDjB0af0g19UniDpn3g.

[5] https://mp.weixin.qq.com/s/LdLEGqjqTXAnkfjgY8Lgow.

[6] Motoki, F., Neto, V. P., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. Public Choice, 1-21.

[7] 王秋菊，陈彦宇.多维视角下智能传播研究的学术图景与发展脉络——基于 CiteSpace 科学知识图谱的可视化分析[J].传媒观察,2022(09).

[8] https://mp.weixin.qq.com/s/E-gSBYbRer4qaVZSHJAUzw.

[9]http://www.xinhuanet.com/tech/20240410/28338f7406354ec6a6824f27e8b18c9a/c.html.